DOCUMENT RESUME

ED 064 388 TM 001 634

AUTHOR Frisbie, David A.; Ebel, Robert L.
TITLE Comparative Reliabilities and Validities of True-False and Multiple Choice Tests.
PUB DATE Apr 72
NOTE 8p.; Paper presented at the annual meeting of AERA (56th, Chicago, Ill., April 3-7, 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Comparative Analysis; Data Collection; High School Students; *Multiple Choice Tests; Natural Sciences; Research; Social Studies; *Test Construction; *Test Reliability; *Test Validity

ABSTRACT

A study designed to compare the reliabilities of multiple choice and true-false tests that were constructed to measure the same objectives was conducted. The impetus for this study came from the research reported by Ebel (1971) on the same topic. Subjects were selected from six public high schools. Three phases of testing were required for instrument development and data gathering. Phase I involved collecting item analysis data for one item conversion method and Phase II was used to try out the true-false items. The final phase of testing included 1018 students responding to eight final test forms. The social studies and natural science multiple choice items employed in this study appeared in a widely used battery of achievement tests. The original 70-item multiple choice tests SM (social studies) and NM (natural science) were each administered to a minimum of 100 subjects. The four true-false test forms were each administered to a minimum of 50 subjects in Phase II. The eight final test forms varied according to subject matter, item conversion method, and item form order. The results of this study support the notion that students respond to more true-false than multiple choice items in a given period of time. However, the data indicate that the multiple choice tests were more reliable though they tended to measure the same thing that the true-false tests measured. (CK)

ED 064388

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Comparative Reliabilities and Validities of True-False and Multiple Choice Tests

David A. Frisbie
Wichita State University

Robert L. Ebel
Michigan State University

TM 001 634

FILMED FROM BEST AVAILABLE COPY

A paper presented at the 56th annual meeting of the American Educational Research Association, Chicago, April 3-7, 1972.

Purpose of the Study

This study was designed to compare the reliabilities of multiple choice and true-false tests that were constructed to measure the same objectives. A second purpose was to determine if multiple choice tests and the true-false tests derived from them measured the same thing.

Background

The impetus for this study came from the research reported by Ebel (1971) on the same topic. His data, in general, supported the notion that true-false tests can be just as reliable as multiple choice tests and both measure relatively the same thing. Two assumptions made in the original study were eliminated in the present study. Data were gathered to determine the ratio of the number of true-false to multiple choice items attempted by examinees in a fixed period of time. This ratio was required to adjust the K-R20's of the true-false tests to equate testing time. The ratio was estimated (2 to 1) in the original study. The second change was to use a systematic and relatively objective procedure for converting test items from multiple choice to true-false form. Two different conversion methods were employed in the present study. In the earlier study development of the true-false items involved considerable subjective judgment on the part of the item writer.

The bulk of the studies reported in the literature that deal with reliability and validity of tests of varying item form were done in the late 1920's and early 1930's when objective examinations began to flourish (Frisbie, 1971).

Sample

The subjects that participated in this study were selected from classrooms in six public high schools in Michigan. Classrooms and schools cooperated on a voluntary basis but were originally approached so that the final sample might represent a cross section of non-urban high school students in science and social studies achievement.

- -

Table 1 Here

- -

Three phases of testing were required for instrument development and data gathering. Phase I involved collecting item analysis data for one item conversion method and Phase II was used to try out the true-false items. The final phase of testing included 1018 students responding to eight final test forms (see Table 1).

Instrumentation

The social studies and natural science multiple choice items employed in this study appeared in a widely used battery of achievement tests. The items were written to measure knowledge and understanding of concepts that are part of the current secondary school curriculum.

The judgmental conversion method (J) required secondary science and social studies teachers to judge the quality of the multiple choice distractors from the items in their respective areas of expertise. They were directed to select the distractor for each item that appeared to be most plausible for making a false statement with the stem. The use of this method resulted in 41 false

and 29 true statements in social studies and 45 false and 25 true statements in natural science. The two 70-item true-false tests were labeled forms SJ (social studies) and NJ (natural science).

The original 70-item multiple choice tests, SM (social studies) and NM (natural science) were each administered to a minimum of 100 subjects in classrooms similar to those involved in the final (Phase III) part of the study. Item analysis data was used to calculate a lower-upper discrimination index for each item response alternative. The foil with the largest lower-upper difference for each item was used to make a false statement with the stem. The discrimination conversion method (D) furnished 37 false and 33 true statements for form SD (social studies) and 37 false and 33 true statements for form ND (natural science).

The four true-false test forms were each administered to a minimum of 50 subjects in Phase II of testing. Three items were slightly revised based on this try-out and all true-false items were then incorporated in eight forms for final testing.

The eight final test forms varied according to subject matter, item conversion method, and item form order. The composition of these forms is indicated by Figure 1. Form SJA, for example, consisted of items 1-35 of the original

- -
Figure 1 Here
- -

multiple choice form (SM) and items 36-70 of form SJ (social studies items converted by the judgmental method). Form SJB was comprised of items 1-35 of form SJ and items 36-70 of form SM.

The four final forms in each subject matter area were administered to randomly selected students in classrooms. Subjects were stopped after eight minutes of testing and were asked to circle the number of the item on which they were working. This data was used to determine the amount of time required to respond to items of each of the two forms.

Results

A K-R_{20} was computed for each of the two subtests in each of the eight final test forms. The reliabilities of the true-false subtests were then adjusted to permit comparison of the two item forms on the basis of equal amounts of testing time, rather than on the basis of equal numbers of items. Since the subjects in this study responded to 25.59 true-false items in eight minutes, but only 17.04 multiple choice items in the same amount of time, the value 1.5 was used for n in the Spearman-Brown formula. The reliability coefficients for the 16 subtests are recorded in Table 2.

- -
Table 2 Here
- -

The difference between reliability coefficients for the subtests using the two item forms (multiple choice vs adjusted true-false) was tested for statistical

significance using a paired-t test. The difference in favor of the multiple choice items was significant beyond the .001 level. No significant difference (p less than .50) was found between the reliabilities of the true-false tests derived by the two different methods of item conversion.

Each subject received a score on the multiple choice and on the true-false test to which he responded. A Pearson product-moment correlation was calculated between subtest scores on each of the eight forms. Table 3 shows the correlation coefficients and the coefficients corrected for attenuation. A t-test statistic

- -
Table 3 Here
- -

developed by Forsyth and Feldt (1969) was used to generate 90% confidence intervals for the eight disattenuated coefficients. The upper and lower limits for these intervals are depicted in Table 4. The hypothesis that the disattenuated correlation coefficient does not differ from unity is supported in six of the eight cases.

- -
Table 4 Here
- -

Conclusions and Discussion

The results of this study support the notion that students respond to more true-false than multiple choice items in a given period of time. However, the data indicate that the multiple choice tests were more reliable though they tended to measure the same thing that the true-false tests measured. These generalizations require some cautionary remarks.

The original tests used in this study were not typical of those constructed by classroom teachers. The items were cast to measure primarily understandings and relationships between concepts. The reliabilities of these tests were much higher (.90) than the reliabilities classroom teachers achieve with their instruments. It is possible that the results of this study would be different if a typical teacher-made multiple choice test had been used originally. The shorter test with less discriminating items would probably yield a smaller range of scores and, therefore, smaller reliability coefficients.

The concurrent validity data should be interpreted with some care. Two of the eight confidence intervals failed to include unity whereas two of the eight (NJB and NDB) were almost certain to include unity by inspection. The probability that all eight confidence intervals included the true population value of the corrected coefficient was 0.43.

The relatively large estimated standard errors of the disattenuated correlation coefficients (see Table 4) caused several of the confidence intervals to be relatively wide. These large estimates were a function of half-test reliabilities for the multiple choice and true-false tests. The median half-test reliabilities were .730 and .431 for multiple choice and true-false tests, respectively

Though the data from the confidence intervals support the hypothesis that true-false and multiple choice tests measure the same thing, the data are not conclusive. The variability in observed correlation coefficients (see Table 3) may be explained in terms of sampling fluctuations, yet these may not account for all of the discrepancies.

It may be true that multiple choice and true-false tests require somewhat different abilities of the examinees. For example, a student may mark a statement true because he could not think of a counterexample, a situation or occurrence that would make the proposition false. His search for a counterexample may have been bounded by time limits or the length to which he could stretch his mind or the depth of his retrieval system that he could penetrate. The multiple choice item, however, limits the universe of comparisons that the individual must make. He can decide which alternative makes a true statement with the item stem and then review the remaining alternatives to determine if any of them is a counter example for the true statement. Though individuals probably differ in the responding schemes they use, their manners of responding to true-false and multiple choice items may depend on somewhat different abilities. The observed correlation coefficients in this study may reflect these differences. The question then arises, if the two item types measure different things, which one best measures what we want to measure? If we are satisfied that our achievement test measures relevant tasks, what suitable external criterion could be used for prediction? When that suitable criterion is discovered we will probably use it to measure achievement instead of our multiple choice or true-false test.

The data from this study do not provide support for those individuals who believe that true-false items are as effective as multiple choice items for measuring classroom achievement. Though the longer true-false tests were less reliable, they exhibited a potential for more adequately sampling the domain of social studies and natural science than did the multiple choice tests. Students could theoretically, attempt 105 true-false items in the time required to respond to 70 multiple choice items, though the former test may be somewhat less reliable.

Though the ratio remained constant (1.5), students attempted slightly fewer natural science than social studies items in eight minutes of testing. This suggests that no hard and fast rules can be formulated regarding the amount of time required to respond to different types of items without considering item content as well.

References

1. Ebel, R. L. "The Comparative Effectiveness of True-False and Multiple Choice Achievement Test Items." Paper presented at the American Educational Research Association Annual Meeting, New York City, February, 1971.

2. Forsyth, R. A. and Feldt, L.S. "An Investigation of Empirical Sampling Distributions of Correlation Coefficients Corrected for Attenuation". Educational and Psychological Measurement, XXIX (Spring, 1969), pp. 61-71.

3. Frisbie, D. A. "Comparative Reliabilities and Validities of True-False and Multiple Choice Tests". Ph.D. Thesis, Michigan State University, 1971.

TABLE 1

Sample Used in Three Phases of Testing

GRADE		PHASE I	PHASE II	PHASE III
9	Social Studies	0	0	0
	Natural Science	49	24	180
10	Social Studies	0	42	145
	Natural Science	27	47	141
11	Social Studies	72	42	104
	Natural Science	18	35	129
12	Social Studies	30	17	260
	Natural Science	7	0	59
	Total	203	207	1018

FIGURE 1

Arrangement of Test Forms Used in Phase III

Test Form	Subtest order	
SJA	MC	TF
SJB	TF	MC
SDA	MC	TF
SDB	TF	MC
NJA	MC	TF
NJB	TF	MC
NDA	MC	TF
NDB	TF	MC

TABLE 2

K-R_{20} Reliabilities for Final Subtest Forms

	Subtest		
	Multiple	True-False	
Test Form	Choice	Original	Adjusted
SJA	.796	.708	.785
SJB	.827	.654	.739
SDA	.805	.498	.598
SDB	.851	.641	.728
NJA	.835	.759	.825
NJB	.852	.612	.703
NDA	.854	.704	.781
NDB	.862	.645	.732

TABLE 3

Correlation Coefficients for Multiple Choice and True-False Subtest Scores on Each Final Test Form

Test Form	r_{mt}	$r_{\infty\infty}$	N
SJA	.578	.769	126
SJB	.697	.947	127
SDA	.564	.891	128
SDB	.430	.582	128
NJA	.661	.831	126
NJB	.728	1.009	129
NDA	.710	.916	125
NDB	.825	1.107	129

TABLE 4

Confidence Intervals for Disattenuated MC-TF Correlation Coefficients

Test Form	Est. Standard Error	Upper Limit*	Lower Limit*
SJA	.0710	.937	.601
SJB	.0629	1.051	.844
SDA	.1275	1.090	.682
SDB	.0913	.733	.431
NJA	.1850	1.135	.527
NJB	.2075	1.350	.668
NDA	.1741	1.202	.620
NDB	.1750	1.395	.819

*90% confidence intervals